

The *Drosophila melanogaster* Actin 5C Gene Uses Two Transcription Initiation Sites and Three Polyadenylation Sites To Express Multiple mRNA Species

BEVERLEY J. BOND¹ AND NORMAN DAVIDSON^{2*}

Division of Biology,¹ and Department of Chemistry,² California Institute of Technology, Pasadena, California 91125

Received 23 December 1985/Accepted 10 March 1986

At least six mRNAs are made from the *Drosophila melanogaster act5C* gene. We investigated the structures of these RNAs in detail and determined that they are heterogeneous at both their 5' and 3' ends. At the 5' end there were two nonhomologous leader exons which were alternately spliced to the remainder of the gene. These leader exons mapped to 1.7 and 0.7 kilobases, respectively, upstream of a common splice acceptor site which was eight base pairs 5' to the translation initiator AUG. Exon 1 is 147 bases in length, while exon 2 is 111 bases. A consensus TATA sequence was found roughly 30 base pairs upstream from exon 1, but none was found in the analogous position upstream of exon 2. The transcript length diversity arose principally from the use of three polyadenylation sites. This gave rise to RNA molecules with 3'-untranslated regions of roughly 375, 655, and 945 base pairs. With two start sites and three termination sites, this gene has the potential to produce six different transcripts. All six possible transcripts were present in whole fly mRNA. Transcripts containing the two different leader exons were found in roughly the same relative quantities through development. In contrast, the various 3' ends were differentially represented through development.

There are six members of the *Drosophila melanogaster* actin multigene family (18, 19, 53). They map to six widely dispersed chromosomal sites. The genes can be placed into three classes that show different developmental patterns of expression (20). Two of the genes (*act79B* and *act88F*) are expressed in the late pupae stage and in adults and encode the major fibrillar and tubular adult-specific muscle isoforms, including those for the flight and jump muscles. *act57A* and *act87E* are expressed in larval and late pupal-adult stages. They are believed to encode actins for larval musculature and abdominal muscles of the adult. *act5C* and *act42A* encode the cytoskeletal actins of nonmuscle cells. They are expressed in the whole animal in most stages of development and are the only actins expressed in early embryos.

Insofar as they have been mapped, intron positions and sequences of the six genes are not conserved, with the exception of the intervening sequence at codon 307 which is present in the *act88F* and *act79B* genes (18, 46). However, the protein-coding regions are rather highly conserved, and there is 85 to 95% amino acid sequence homology among the six actin isoforms.

In this study we were concerned with the structure of the *act5C* gene. The structure of the gene, as it was known prior to this study, is depicted in Fig. 1. R-loop studies indicated that there is a short exon (exon 1) about 1.7 kilobases (kb) upstream of the main exon (exon 3). Sequence data suggest that the protein-coding region is entirely contained in exon 3 and that there is a consensus splice acceptor sequence 8 nucleotides upstream of the translation initiator ATG (18). Sequence data and results of in vitro transcription studies provided a tentative identification of a cap site at the 5' end of exon 1 as well as a TATA box and an activating sequence upstream (40; D. Price, B. Korber, J. Topol, and C. S. Parker, personal communication).

RNA gel blots showed that there are three *act5C* tran-

scripts with molecular lengths of 2.2, 1.95, and 1.7 kb (20). Their relative intensity in whole animal RNA varies with developmental stage. In general, the 1.95-kb band is the most intense. We hypothesized that the reason for the length difference is alternative choices of polyadenylation sites in the 3'-untranslated region.

This study was undertaken to obtain decisive information about the structures of the several transcripts. The results reported below show that in addition to the mRNA of structure exon 1-exon 3, as depicted in Fig. 1, there is an additional cap site (and presumed transcription start site) between exons 1-3 and an additional mRNA of structure exon 2 and exon 3, as shown in Fig. 1. Exons 1 and 2 are not greatly different in length. The main cause of the three length classes is the existence of three alternate polyadenylation sites. Our data indicate that in whole animal RNA, exons 1 and 2 are used with approximately equal probability in all stages of development and with all three polyadenylation sites. Furthermore, there is some developmental variability in the usage of the three polyadenylation sites.

MATERIALS AND METHODS

Preparation of RNA. RNA was prepared by a modified guanidinium thiocyanate method (12, 19). Typically, about 15 ml of guanidinium solution was used for each gram of tissue. After the addition of CsCl, the mixture was centrifuged for 15 min at $10,000 \times g$ in an SS34 rotor. The cuticles and other debris formed a band on top of the CsCl solution which was discarded. The solution was then spun in the ultracentrifuge as described above. Poly(A)⁺ RNA was selected by oligo(dT) chromatography (2).

Primer extension. A synthetic 24-base oligonucleotide that was complementary to the actin 5C RNA between codons 26 and 34 was used as a primer. The sequence of this primer was 5'-TCGATGGGAAGACGGCGCGGGGAG-3'. It was synthesized by S. Horvath (California Institute of Technology) on an automated DNA synthesizer (29) and then

* Corresponding author.

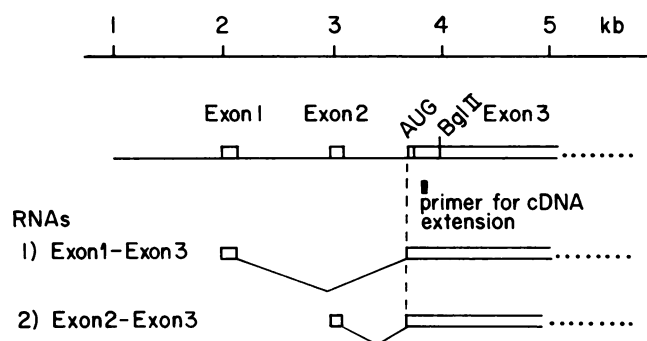


FIG. 1. Transcript map of the 5' end of the *act5C* gene. Exon sequences are represented by open boxes, while intron and flanking sequences are solid lines. The alternate splicing pathway is depicted below. The location of the primer used for primer extension experiments is also indicated.

purified from a 20% polyacrylamide-8 M urea gel and precipitated with ethanol. It was 5' end labeled with polynucleotide kinase and [32 P]ATP (~7000 Ci/mmol). Labeled primer (150 pmol) and poly(A) $^{+}$ RNA (1 μ g) were suspended in 20 μ l of 70% formamide containing 0.4 M NaCl, 40 mM piperazine-*N,N'*-bis(2-ethanesulfonic acid), and 1 mM EDTA. After incubation for 2 min at 60°C and 2 h at 37°C, the reaction mixture was diluted to 200 μ l. Ammonium acetate (200 μ l) and isopropyl alcohol (240 μ l) were added. After 15 min at room temperature, the nucleic acids were recovered by centrifugation. The RNA precipitation procedure removed most of the excess primer molecules. The ethanol-rinsed and dried pellet was suspended in 15 μ l of a 2 \times cDNA mixture (0.1 M Tris [pH 8.3]; 12 mM magnesium acetate; 0.12 M NaCl; 20 mM dithiothreitol; 2 mM each of dATP, dGTP, and TTP). To this was added 2 μ l of 10 mM dCTP, 1.9 μ l of 0.8 μ g of actinomycin D per ml, 15 U of reverse transcriptase, and water to 30 μ l. The solution was incubated at 37°C for 45 min and precipitated with ethanol. The precipitated products were run on alkaline agarose gels (36) or, after RNase A digestion of the RNA templates, on 6% polyacrylamide-8 M urea gels.

S1 nuclease and exonuclease VII mapping of transcripts. Typically, 10 ng of 32 P-labeled DNA fragments was hybridized to 1 μ g of poly(A) $^{+}$ RNA in 80% formamide at 55°C and then digested with S1 nuclease (6, 15) or exonuclease VII (7). For the exonuclease VII reactions, after the hybridization step 1 U of enzyme was added to each reaction in 100 μ l of cold exonuclease VII buffer (30 mM KCl, 10 mM Tris [pH 7.4], 10 mM EDTA). After incubation at 45°C for 1 h, the reactions were precipitated with ethanol, with yeast tRNA used as a carrier. Ethanol-washed and dried samples were loaded on alkaline agarose gels (36). Protected fragments were detected by autoradiography of the dried gels.

Isolation of cDNA clones. A *D. melanogaster* embryo (8- to 20-h old) cDNA library in the vector lambda *gt*10 (kindly provided by L. Kauvar) was screened by the plaque lift technique of Benton and Davis (4) with nick-translated probes.

DNA sequencing. DNA sequencing was done by the chemical modification technique (35).

RNA blotting. Poly(A) $^{+}$ RNAs were electrophoresed on formaldehyde-agarose gels and blotted to nitrocellulose by a modification of the technique of Thomas (52). Blotting, hybridizations, and blot washings were performed as described previously (20). Hybridization probes were gel isolated and labeled by nick translation.

RESULTS

The *act5C* gene has two transcription start sites. The structure of the 5' end(s) of the transcripts was first investigated by primer extension experiments. The primer was a synthetic oligonucleotide which was chosen to fulfill two criteria. First, it had to be able to prime cDNA synthesis from all *act5C* transcripts, regardless of their 5'-untranslated sequences. For this reason, the sequence was taken from the protein-coding portion of the gene. Second, it had to be specific for the *act5C* gene and not hybridize to other actin mRNAs. The actin protein-coding regions, in general, were very homologous, but a search of the protein-coding sequences revealed a 24-nucleotide stretch which contained six mismatches between the *act42A* and *act5C* genes. This sequence, which starts 80 base pairs (bp) downstream from the *act5C* gene initiator AUG, was used for the primer, and early embryo RNA was used for the experiments because only these two actin genes were expressed at that developmental stage. Hybridizations were done under conditions in which the primer would hybridize only to the *act5C* RNA.

Surprisingly, the primer extension experiments showed four bands with lengths of 258, 255, 222, and 213 nucleotides, and they were of approximately equal intensity (Fig. 2). These data indicate that transcripts of the gene have at least two and perhaps as many as four different 5' ends but do not give any information about where they map on the genomic DNA.

To map the location of the 5'-untranslated sequences with respect to the genomic DNA, S1 nuclease and exonuclease VII experiments were done. A 4.2-kb kinase-labeled *Bgl*II fragment of genomic *act5C* DNA (Fig. 3B) was used as a hybridization probe to poly(A) $^{+}$ RNA from different developmental stages. This DNA fragment contains 250 bp of actin protein-coding sequence and 4 kb of upstream sequences.

Hybrids formed between early and late embryo poly(A) $^{+}$ RNA and the labeled DNA fragment were digested with exonuclease VII or S1 nuclease. The exonuclease VII- and S1 nuclease-protected bands are displayed on an alkaline agarose gel (Fig. 3). The two enzymes protected fragments with different lengths, indicating that splicing occurs at the 5' end of the gene. The S1 nuclease-protected band was roughly 300 bases in length, indicating the presence of a splice acceptor site just upstream of the AUG codon in all *act5C* transcripts. Exonuclease VII digestion yielded two fragments with lengths of 0.95 and 1.98 kb which were of roughly equal intensity.

These results indicate that exon sequences terminating at their respective 5'-end cap sites are present in the genomic DNA at roughly 0.7 kb (exon 2) and 1.7 kb (exon 1) upstream from the splice acceptor site. The location of exon 1 had previously been found by R-loop analysis (17) and by in vitro transcription experiments (D. Price and C. Parker, personal communication). Exon 2 was not found by R-loop experiments.

These results do not determine the size or number of the upstream exons in the several transcripts. Furthermore, the relationship between the two exonuclease VII digestion products and the primer extension products was not resolved by this experiment.

Sequencing of the 5' leader exons. To determine the size and number of upstream exons, we isolated and analyzed cDNA clones for the *act5C* gene. A *D. melanogaster* (8- to 20-h-old embryo cDNA library was obtained from Larry Kauvar. It was screened with a genomic probe which

contained about 4 kb of sequences upstream of the AUG codon and no protein-coding sequences. Actin protein-coding sequences are highly homologous, but the untranslated sequences are not (18, 19). This probe was chosen to maximize the chances of getting *act5C* specific full-length cDNA clones.

A screen of 2×10^4 phage gave 16 positive clones. The 5' ends of 10 of these clones were sequenced from the 5' ends of the inserts to a *Sa*I site 30 bases inside of the coding region of the gene. We sequenced into the coding region of each cDNA clone to be sure that we had isolated *act5C* clones. There were enough base substitutions in the first 30 bp of the protein-coding sequence to identify unambiguously each actin gene. We also sequenced a 900-bp region of *act5C* genomic DNA that was shown by exonuclease VII experiments to contain exon 2. The genomic sequence data and their interpretation, as deduced from comparison with the

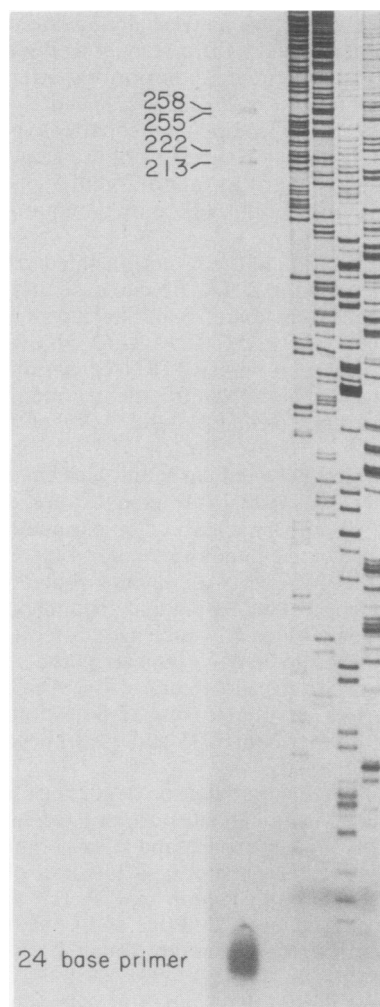


FIG. 2. Mapping of the 5' ends of *act5C* RNAs by primer extension. A 24-base synthetic oligonucleotide homologous to the *act5C* gene between codons 26 and 33 (Fig. 1) was labeled with kinase and hybridized to 2 μ g of poly(A)⁺ RNA from 0- to 4 h-old embryos. cDNA synthesis was carried out as described in the text. The extended products were run on a 6% sequencing gel. A dideoxy-sequencing ladder is shown in the four rightmost lanes which was used for size comparison. Numbers to the left of the gel are in bases.

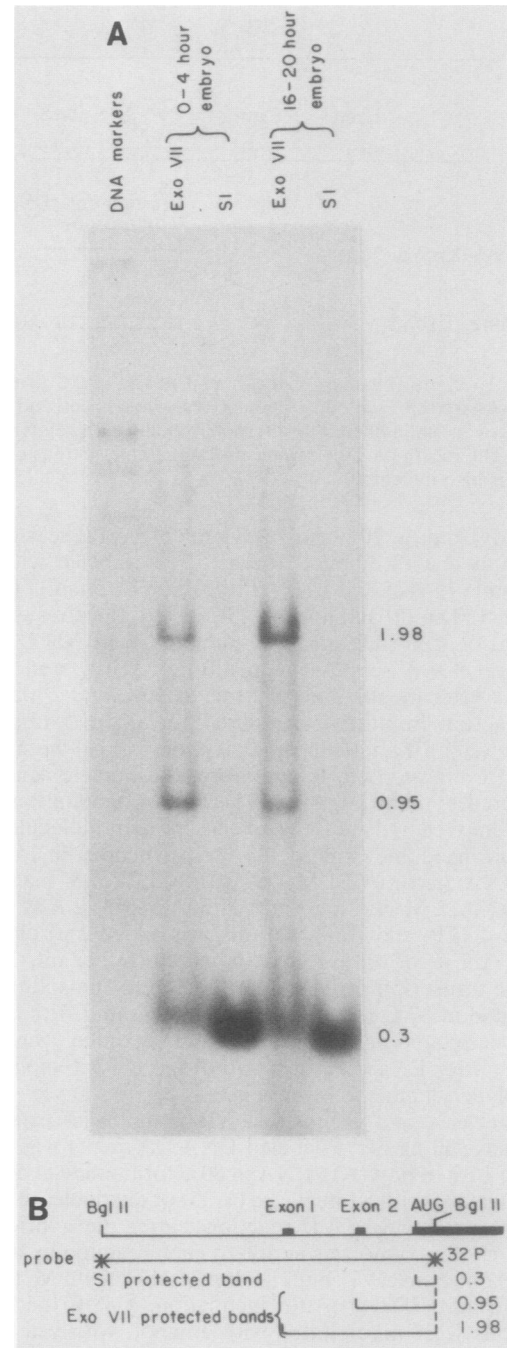


FIG. 3. (A) S1 nuclease and exonuclease VII mapping of the 5' end of the gene. Poly(A)⁺ RNA (1 μ g) from 0- to 4-h-old or 16- to 20-h-old embryos was hybridized to a kinase-labeled *Bgl*II restriction fragment. The hybrid molecules were digested with either S1 nuclease or exonuclease VII. The protected products were run on a 1% alkaline agarose gel as shown. The left-hand lane contains a *Hind*III-*Eco*RI digest of lambda DNA which was used as a size marker. Numbers to the right of the gel are in kilobases. (B) The probe fragment and protected bands.

cDNA sequences (see below), are shown in Fig. 4. A 430-bp region of genomic DNA surrounding exon 1 has been sequenced previously (40; D. Price, personal communication).

All of the cDNA clones had the same splice acceptor site which was 8 nucleotides upstream of the AUG, as previously

```

-260 ..... ..tggaagt acactcttca
-200 tggcgatata caagacacac acaagcacga acaccagtt gcggaggaaa ttctccgtaa
-140 atgaaaaccc aatcggcgaa caattcatat ccatatatgg taaaagtgtt gaacgcgact
-80 tgagagcgga gagcattgcg gctgataagg ttttagcgct aagcgggctt tataaacgg
-20 gctgcgggac cagttttcat ATCACTACCG TTTGAGTTCT TGTGCTGTGT GGATACTCCT
 41 CCCGACACAA AGCCGCTCCA TCAGCCAGCA GTCGTCTAAT CCAGAGACAC CAAACCGAAA
101 GACTTAATTT ATATTTATTT AATTAATTTT AATAAAACAC ACCAAATgta agtagctttc
161 cccttcccaa caacaaaaca ccatcgaacc actcccacca agaaaaagca ata.....

591 ..... aatgtacata catacagtat atgcatatta taatctgtaa aactagatca
651 gggtcttgaa aatagtgacg taggcagccg ttttggctga agcagaaatt tttgccgggt
711 tttcaaagt gtagttgcaa aaatggagaa aaccttcgag cattcgttca tatacacaca
771 ctcacgcgca aaataacgag agagagtgtg tgtgtgtgtg agagagcgaa agccagacga
831 cggtttgctt ttcgcctcga aacatgacca tatatggtca caaaacttgg ccgccgcaat
891 tcaacacacc agcgtcttcc ttcgcacca tagcgaccat gcggcggagc gagcgagatg
951 gcgagagcga gcgacgccta tggcgacgtc gacgcaggca gcgattgaaa aacgcagtta
1011 actggcattc aacattcacc agccactttc AGTCGGTTTA TTCCAGTCAT TCCTTTCAAA
1071 CCGTGCGGTC GCTTAGCTCA GCCTCGCCAC TTGCGTTTAC AGTAGTTTTC ACGCCTTGAA
1131 TTTGTTAAAT CGAACAAAAA Ggtaaagttt aactagcttt gaaaagtttc gtggctctta
1191 attgttaaatt tttctagagt gcgttttagtg tttttttttt tttttatttt gtaatgttaa
1251 tttcgggttc caattcgagt ttttaggcagc cgcactttta agggcgcata cacacaggca
1311 actgtgctct ctttgcggtt tttttttgca cgggcattcg ttaagtgtcg tctagaagct
1371 tctcccttcc .....

1681 ..... ggtaacaaaa aactaatggg aaatccgcat tctttccatt gcagCTTACA

1741 AA Met Cys Asp Glu Glu Val Ala Ala Leu Val Val Asp
ATG TGT GAC GAA GAA GTT GCT GCT CTG GTT GTC GAC

```

FIG. 4. DNA sequence of exon 1, exon 2, intron, and 5'-flanking sequences of the *act5C* gene. The probable 5' terminus of exon 1 is designated as nucleotide 1. Capital letters indicate exon sequences. Lower case letters are intron and flanking sequences. TATAA (exon 1) and TTAA (exon 2) sequences just preceding the transcription starts are underlined. Gaps in the sequence occur between nucleotides 213 and 601 and again between nucleotides 1381 and 1691. The sequence of the protein-coding region is shown only up to the *SalI* site which was the 3' termination point of the cDNA clone sequencing. The sequence between -217 and 213 was determined by D. Price. The rest was done by sequencing of both strands by the method of Maxam and Gilbert (35).

surmised (20), and they all contained *act5C* protein-coding sequences. The clones fell into two classes with respect to sequences further in the 5' direction.

Of the 10 clones, 7 had roughly 147 bp of sequence joined directly to the previously mentioned splice acceptor site. A comparison of these sequences with known genomic sequences revealed that they are completely homologous, with a 147-bp stretch of sequence about 1.7 kb upstream of the initiator AUG (exon 1). A primer extension product made from an RNA of this structure would be 258 bases in length or the same as the longest of the four products seen. Five of the seven clones were full length, while one was 3 bp shorter and the final clone was 25 bp shorter. The clone with the

144-bp exon 1 sequence was of the correct size to correspond to the primer extension product with a length of 255 nucleotides. We do not know whether this is a real alternate start site or a premature stop by reverse transcriptase. We presume that the clone that was 25 bp short is an incomplete cDNA.

The remaining three clones had a different sequence spliced to the acceptor site. This sequence had no homology to a 450-bp region which included exon 1. Rather, it was identical to sequences roughly 0.7 kb upstream of the initiator AUG codon at the site predicted by the exonuclease VII analysis (exon 2). In two of these clones, this exon was 111 bp in length, while the third clone was 9 bp shorter,

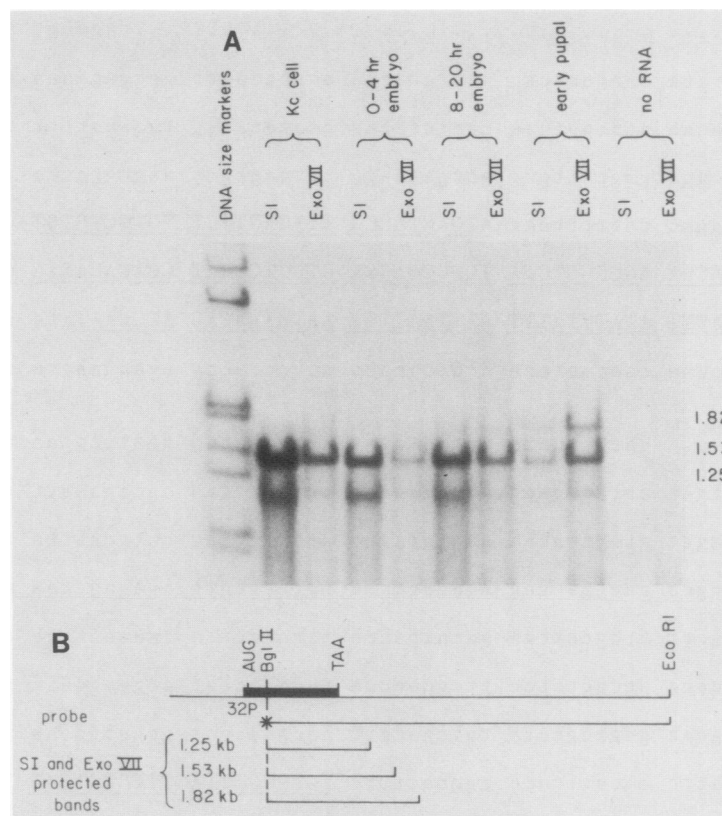


FIG. 5. (A) S1 nuclease and exonuclease VII mapping of the 3' end of the gene. Poly(A)⁺ RNA (1 μ g) from the developmental stages indicated was hybridized to a 4-kb *Bgl*II-*Eco*RI fragment labeled with the Klenow fragment of DNA polymerase I. The hybrid molecules were digested with S1 nuclease or exonuclease VII, and the protected products were run on a 1% alkaline agarose gel. Numbers to the right of the gel are in kilobases. (B) The probe and protected fragments. The DNA markers are a lambda *Eco*RI-*Hind*III digest.

corresponding to the observed primer extension bands of length 222 and 213 nucleotides. Again, we do not know whether the shorter product is real or a reverse transcriptase artifact. This kind of microheterogeneity would not have been seen in our lower resolution exonuclease VII experiments.

Of the 10 clones, 4 had an extra guanine nucleotide between the poly(dC) tail used in the cDNA cloning and the transcription start that did not correspond to the genomic sequence. This may be due to reverse transcription of the cap nucleotide.

The correspondence between the cDNA clones and the primer extension products was confirmed by sequencing the latter (data not shown). These results indicate that at least two independent transcription start sites are used to give rise to two classes of actin 5C mRNA. They are alternately spliced to the body of the gene.

Three sites of polyadenylation are used by the *act5C* gene. The difference in length between the two alternate leader exons is not sufficient to be resolved on an RNA blot. The major source of the RNA size heterogeneity seen on RNA blots must lie elsewhere. Alternate splicing and different sites of polyadenylation are two possibilities.

The structure of the 3' ends of the RNAs was investigated by S1 nuclease and exonuclease VII experiments. A 4-kb fragment originating within the protein-coding region (at the same *Bgl*II site used for the 5' experiments) and extending 3 kb beyond the translation stop codon was used as the hybridization probe. This labeled fragment was hybridized to poly(A)⁺ RNA from various developmental stages or from the Kc line of *D. melanogaster* cultured cells. The hybrids

were treated with exonuclease VII or S1 nuclease, and the protected products were run on alkaline agarose gels (Fig. 5).

For each RNA used, the S1 nuclease- and exonuclease VII-protected products were the same length, indicating that no splicing occurs in the transcripts 3' to the *Bgl*II site. In all cases, bands of 1.25 and 1.53 kb were seen. These corresponded to 3'-untranslated regions of 375 and 655 bp, respectively. In the 8- to 20-h-old embryo and early pupal RNAs, a larger 1.82-kb band was also seen. This corresponds to a 3'-untranslated region of 945 bp. These results indicate that different sites of polyadenylation are used to yield three size classes of RNAs. These size differences are of an appropriate size to account for the three different *act5C* species with molecular lengths of 1.7, 1.95, and 2.2 kb, as resolved on RNA blots.

Differential expression of the *act5C* transcripts. It is also apparent from the S1 nuclease and exonuclease VII data and from RNA blots (20) that use of the different sites of polyadenylation is developmentally regulated. The intermediate site of polyadenylation was used in all stages of development tested. Transcripts of this type were generally the most abundant. Transcripts with the shortest 3'-untranslated region were also present at almost every stage tested, but they were usually present at a lower level and peaked at different times during development. Transcripts which used the last polyadenylation site and had the longest 3'-untranslated region were only found at specific times in development. They were most abundant at mid-embryo and early pupal stages.

To study the relative representation of the two different

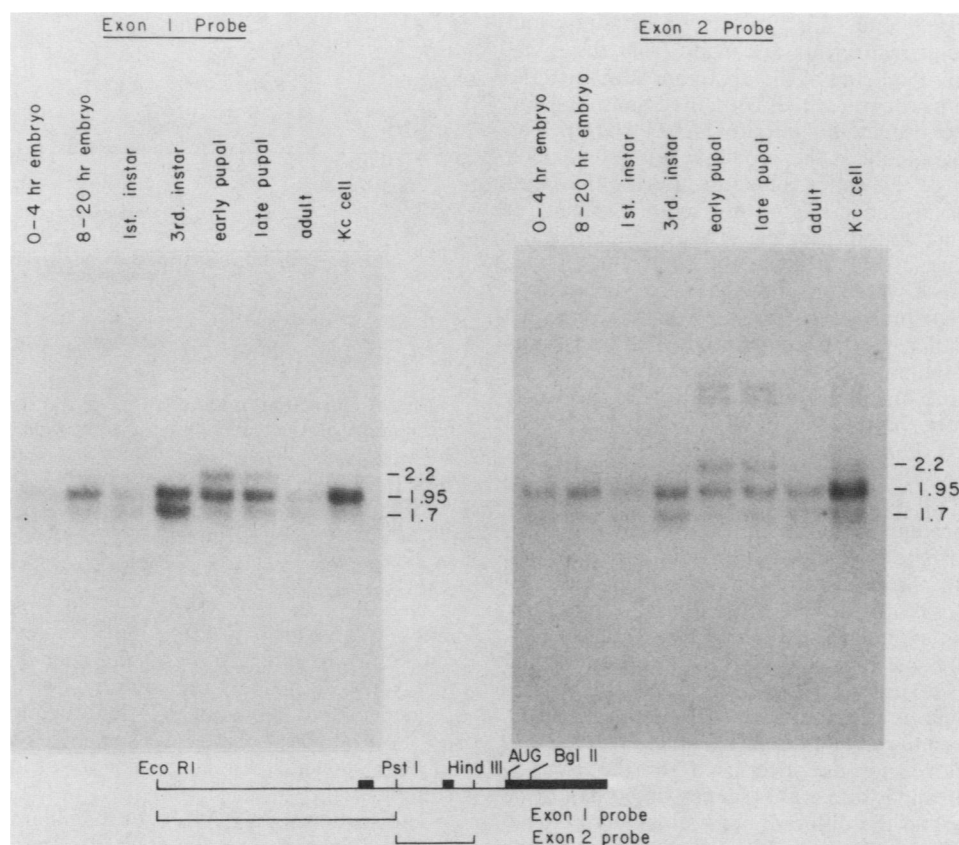


FIG. 6. Developmental RNA blots probed with exon 1- and exon 2-specific probes. Identical blots containing 1 μ g of poly(A)⁺ RNA from the developmental stages indicated, and Kc cells were probed with nick-translated exon 1- or exon 2-specific fragments, which are indicated in the diagram below. Numbers to the right of the gels are in kilobases.

leader exon sequences among *act5C* transcripts, parallel RNA blots were made which contained poly(A)⁺ RNA from several different stages of *D. melanogaster* development and *D. melanogaster* Kc cells. These were probed with exon 1- and exon 2-specific probes (see Fig. 2; Fig. 6).

The blots that were hybridized with exon 1 and 2 probes looked very similar. Both leader exons were used in RNAs of 1.7, 1.95, and 2.2 kb, indicating that all three polyadenylation signals are used with each start site (Fig. 6). Both leader exons were represented in transcripts at each stage of development studied. The level of these transcripts appeared to rise and fall with a similar pattern. No remarkable differences in the pattern of expression of these two leader exons were seen, with the exception of two higher molecular weight species which were seen only in pupal RNA and were homologous only to the exon 2 probe. The identity of these bands is unknown.

DISCUSSION

We have shown that six different transcripts are made from the *D. melanogaster act5C* gene. These results are shown in Fig. 7 and discussed below.

The *act5C* gene has two transcription start sites. We found that two classes of RNA with respect to 5'-untranslated sequences are made from the *act5C* gene. Their 5' sequences were encoded by two distinct leader exons which were located 1.7 and 0.7 kb upstream of the translation initiator AUG. These leader sequences were alternately spliced to a common splice acceptor site which is 8 bp 5' of the AUG

codon. We saw some microheterogeneity in the exact initiation sites of exons 1 and 2, both in primer extension experiments and cDNA clones. It is not clear whether all of these sites are actually used in vivo or if they are reverse transcriptase artifacts.

We did not see any developmental specificity in the use of the two leader exons. Transcripts containing the alternate exons displayed approximately the same pattern of expression through development. It is possible that transcription from the two start sites is regulated but that we would not have detected the difference on the RNA blots that we did. For example, there may be tissue-specific expression of the two *act5C* start sites at a given developmental stage. Alternatively, transcription from the two promoters could be linked to the cell cycle. Further experiments will be necessary to test these possibilities.

Many other cases of alternate transcription start sites and alternate first exons for eucaryotic genes are known. The resulting transcripts are often made in a developmental or tissue-specific manner, as in the following examples. In each case, the structural organization of the transcripts is slightly different.

The mouse α -amylase 1^a gene (25, 26, 48, 57), like *act5C*, has two leader exons which are alternately spliced to a common third exon. The distal promoter is 30 times stronger than the proximal promoter and drives transcription of a parotid gland-specific mRNA. Transcription from the proximal promoter produces an mRNA which accumulates to a level that is 100 times lower in the liver, parotid gland, and pancreas. Two classes of *D. melanogaster* ADH mRNA that

differ only at their 5' ends are transcribed in adults and larvae (5). The adult transcripts are made from the distal promoter and require splicing of the upstream adult-specific leader exon to the next exon which contains 5'-untranslated sequences and the translation initiator AUG codon. The larval transcripts originate at the proximal promoter which lies just upstream of the splice acceptor site for the adult transcript. The larval transcripts do not require splicing at their 5' ends. Transcription from the distal promoter of the yeast invertase gene (10, 43) is under the control of glucose and produces an RNA which encodes the secreted, glycosylated form of the enzyme. Transcription from the proximal start site, on the other hand, is constitutive. The transcripts which originate at the proximal site start within the sequence of the signal peptide of the protein coded for by the larger RNA and therefore produce a protein without a signal sequence, which is therefore intracellular. Neither product requires 5' splicing. The chicken myosin light chain 1 (LC₁) and 3 (LC₃) gene (38) uses alternate splicing of two separate leader exons to produce different transcripts. In this case, the splice acceptor site is in the protein-coding region, and the RNAs code for proteins with different amino-terminal sequences. Transcripts made from the distal promoter encode LC₁ which is expressed in skeletal muscle and heart. LC₃ transcripts are made from the proximal promoter and are expressed in skeletal muscle and gizzard. LC₁ is made earlier in embryonic development (10 to 14 days) than LC₃ which appears in embryos at 14 to 15 days.

Multiple initiation sites are also used for the chicken ovomucoid (21, 30) and lysozyme (24) genes. In both of these cases, however, while the different start sites may be used with different efficiencies, there does not appear to be developmental or tissue-specific use of the start sites.

Many genes have been shown to have microheterogeneity around the cap sites. In these cases several mRNAs are made from a single gene, but they differ by only a few bases in length. Some examples are the human alpha actin gene (27), the *D. melanogaster* myosin LC-2 gene (41), the adenovirus type 2 EII gene (1), the simian virus 40 late genes (16, 22, 23), the polyomavirus late genes (17), the chicken ovalbumin gene (34), the chicken lysozyme gene (24), and the yeast Adh-I gene (3).

In conclusion, this review of the literature indicates that for most of the known cases of alternate transcription start sites and alternate first exons, there is some tissue or stage specificity of expression, a resulting difference in the gene product, or both. In contrast, there is no indication of such differences for the two transcription start sites of the actin 5C gene.

Exon 2 lacks a TATAA sequence. An examination of the DNA sequences immediately 5' of the two cap sites showed that some sequences present in this region in many eucaryotic genes are lacking. We did find a TATAA sequence 30 bp upstream of the transcription start at exon 1, but no such sequence was found in the vicinity of exon 2. An AATT sequence was found 30 bp upstream of the exon 2 cap site. Neither start site was preceded by a CCAAT sequence which is generally found roughly 65 bp upstream from the transcription start.

There are several cases of eucaryotic genes with no TATAA sequence in the region 30 bases upstream of the transcription start. Some examples are the *D. melanogaster* myosin alkali LC (14), the human antithrombin III gene (44), the simian virus 40 late genes (16, 45), and the adenovirus type 2 EII gene (1). Several of the genes mentioned earlier, namely the *D. melanogaster* Adh-I adult promoter (5), the

Act -5C

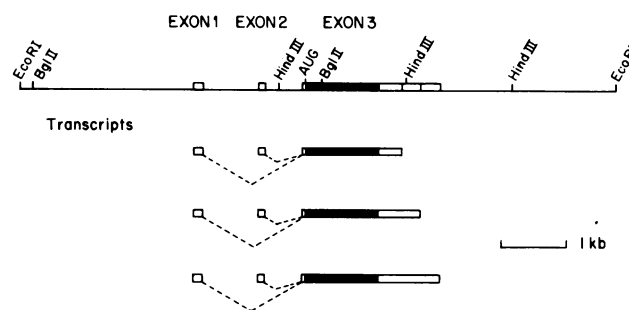


FIG. 7. Structure of the *act5C* gene transcripts. Schematic representation of the generation of transcripts from the *D. melanogaster act5C* gene. The solid boxes represent protein-coding sequences, while the open boxes represent transcribed, untranslated sequences. Intron and flanking sequences are designated by the solid lines.

chicken myosin LC (38), chicken ovomucoid (21, 30), chicken lysozyme (24), *D. melanogaster* myosin LC-2 (41), and yeast Adh-I (3), are missing the consensus TATAA sequences but have an AT-rich region in the analogous position. The lack of TATAA sequences has been postulated to be responsible for heterogeneity around the cap site in some cases.

The sequence ATCAGTC or a similar sequence has been found at the transcription start of several *D. melanogaster* mRNAs (51). The sequences ATCACTA and TTCAGTC were found at the starts of exons 1 and 2, respectively.

Do exons 1 and 2 have independent promoters? Two lines of evidence suggest that the sequences at and upstream of exon 1 are sufficient for the independent initiation of transcription. Price and co-workers (Parker et al., personal communication) have tested a DNA segment containing exon 1 and upstream sequences, but not exon 2, in an in vitro transcription system. They found a high level of transcription initiation at the same site that we mapped as the cap site for exon 1. Furthermore, we made fusions between putative actin promoter sequences and the bacterial chloramphenicol acetyltransferase gene and tested these constructs by transient assays in *D. melanogaster* Kc cells (unpublished data). We found that sequences upstream of exon 1 are sufficient to cause a high level of expression from this promoter in these cells.

However, at present, there is no evidence as to whether transcription can initiate at the cap site of exon 2 under the control of promoter sequences immediately upstream, or whether sequences around the promoter for exon 1 are necessary for production of the RNA beginning at exon 2. Experiments to test this question are in progress.

Multiple sites of polyadenylation. Three major size classes of RNA exist for the *act5C* gene. We found that the major source of this size heterogeneity lies at the 3' end of the gene at which three sites of polyadenylation are used. The first two sites were used in all stages of development studied, but the corresponding mRNAs were not always present in the same relative amounts. The transcripts which terminated after the second polyadenylation signal were always the most abundant. The third and distant site was used only at particular times in development, especially at mid-embryo and early pupal stages. There is as yet no known functional

significance to the different developmental choices of polyadenylation sites in the *act5C* gene.

Most eucaryotic genes have a sequence AATAAA that is roughly 30 bp upstream of the site of polyadenylation in the RNA. We did not determine the precise sites of poly(A) addition for these transcripts. We estimate from S1 nuclease and exonuclease VII analyses that the sites of poly(A) addition lie roughly 375, 655, and 945 bases from the translation stop codon. There are no AATAAA sequences in the 3'-untranslated region of the *act5C* gene (D. Price, E. A. Fyrberg, and B. J. Bond, unpublished data). However, there are closely related sequences 20 to 60 bp upstream from each presumed poly(A) addition site. That is, AATATA and AATGAAA sequences are found roughly 55 and 35 bp, respectively, in front of the proximal poly(A) site. The AATATA sequence is a minor variant that has been seen in other eucaryotic genes, while the AATGAAA variant has been found by Wickens and Stephenson (56) to be an inefficient substrate for poly(A) addition. TATAAA and AATCAAA sequences lie roughly 60 and 20 bp 5' to the middle poly(A) site. The TATAAA is another minor variant in eucaryotes, while the AATCAAA has not been seen previously. The distal site of polyadenylation is preceded by an AATTAAA sequence which lies roughly 45 bp upstream. This variant has been seen in 12% of the RNAs compiled (8).

Many eucaryotic genes generate multiple transcripts from a single gene through the use of different sites of polyadenylation. Some examples of this are the mouse α -amylase gene (54), the dihydrofolate reductase gene (49, 50), the bovine prolactin gene (47), the ovalbumin gene (33), the ovalbumin X and Y genes (28, 32), the human β -tubulin gene (31), the chicken vimentin gene (9, 58), the chicken ovomucoid gene (21), the immunoglobulin heavy-chain genes (11, 13, 37), the major late adenovirus type 2 transcription unit (39), the yeast Adh gene (3), the α -2 microglobulin gene (55), the β -2 microglobulin gene (42), and the *D. melanogaster* myosin alkali LC gene (14).

In many cases the alternative sites of polyadenylation are used with different frequencies, and in the case of the chicken vimentin gene (9) there is some tissue specificity in the use of the different sites. The significance of having multiple mRNA species transcribed from a single gene and differing only in the lengths of their 3'-untranslated sequences, however, is in general unknown.

The only case in which the alternate use of polyadenylation sites has a known functional significance is the case of the immunoglobulin heavy-chain genes. Early et al. (13) showed that the mRNAs for the secreted and membrane-bound forms of immunoglobulin M are transcribed from the same gene and differ only at their 3' ends. A similar arrangement has been found for all of the immunoglobulin heavy-chain classes. Milcarek and Hall (37) showed that the ratio of the membrane-bound and secreted forms of the γ -2b mRNA in the cell is determined by selective use of the alternate polyadenylation sites.

ACKNOWLEDGMENTS

We are grateful to Bette Korber, David Price, Joanne Topol, and Carl Parker for communicating results prior to publication and to William Mattox and Eric Fyrberg for helpful discussions. We thank Larry Kauvar for providing the *D. melanogaster* (8- to 20-h old cDNA library in lambda *gt*10 and Mickey Chien-Tsung Hu for running a primer extension sample on a sequencing gel.

This work was supported by a Public Health Service research grant from the National Institutes of Health.

LITERATURE CITED

1. Baker, C. C., J. Herisse, G. Courtois, F. Gailbert, and E. Ziff. 1979. Messenger RNA for the Ad2 DNA binding protein: DNA sequences encoding the first leader and heterogeneity at the mRNA 5' end. *Cell* 18:569-580.
2. Bantle, J. A., I. H. Maxwell, and W. E. Hahn. 1976. Specificity of oligo(dT) cellulose chromatography in the isolation of polyadenylated RNA. *Anal. Biochem.* 72:413-427.
3. Bennetzen, J. L., and B. D. Hall. 1982. The primary structure of the *Saccharomyces cerevisiae* gene for alcohol dehydrogenase I. *J. Biol. Chem.* 257:3018-3025.
4. Benton, W. D., and R. W. Davis. 1977. Screening (lambda)gt recombinant clones by hybridization of single plaques in situ. *Science* 196:180-182.
5. Benyajati, C., N. Spoerel, H. Haymerle, and M. Ashburner. 1983. The messenger RNA for alcohol dehydrogenase in *Drosophila melanogaster* differs in its 5' end in different developmental stages. *Cell* 33:125-133.
6. Berk, A., and P. Sharp. 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease digested hybrids. *Cell* 12:721-732.
7. Berk, A. J., and P. A. Sharp. 1978. Spliced early mRNAs of simian virus 40. *Proc. Natl. Acad. Sci. USA* 75:1274-1278.
8. Birnstiel, M. L., M. Busslinger, and K. Strub. 1985. Transcription termination and 3' processing: the end is in site. *Cell* 41:349-359.
9. Capetanaki, Y. G., J. Ngai, C. N. Flytzanis, and E. Lazarides. 1983. Tissue-specific expression of two mRNA species transcribed from a single vimentin gene. *Cell* 35:411-420.
10. Carlson, M., and D. Botstein. 1982. Two differentially regulated mRNAs with different 5' ends encode secreted and intracellular forms of yeast invertase. *Cell* 28:145-154.
11. Cheng, H., F. R. Blattner, L. Fitzmaurice, J. F. Mushinski, and P. W. Tucker. 1982. Structure of genes for membrane and secreted murine IgD heavy chains. *Nature (London)* 296:410-415.
12. Chirgwin, J. M., A. E. Przybyla, R. J. MacDonald, and W. J. Rutter. 1979. Isolation of Biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 18:5294-5299.
13. Early, P., J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall, and L. Hood. 1980. Two mRNAs can be produced from a single immunoglobulin (mu) gene by alternative RNA processing pathways. *Cell* 20:313-319.
14. Falkenthal, S., V. P. Parker, and N. Davidson. 1985. Developmental variations in the splicing pattern of transcripts from the *Drosophila* gene encoding myosin alkali light chain result in different carboxy-terminal amino acid sequences. *Proc. Natl. Acad. Sci. USA* 82:449-453.
15. Favaloro, J., R. Treisman, and R. Kamen. 1980. Transcription maps of polyoma virus-specific RNA: analysis by two dimensional nuclease S1 gel mapping. *Methods Enzymol.* 65:718-749.
16. Fiers, W., R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, and M. Ysebaert. 1978. Complete nucleotide sequence of SV40 DNA. *Nature (London)* 273:113-120.
17. Flavell, A. J., A. Cowie, S. Legon, and R. Kamen. 1979. Multiple 5' terminal cap structures in late polyoma virus RNA. *Cell* 16:357-371.
18. Fyrberg, E. A., B. J. Bond, N. D. Hershey, K. S. Mixter, and N. Davidson. 1981. The actin genes of *Drosophila*: protein coding regions are highly conserved but intron positions are not. *Cell* 24:107-116.
19. Fyrberg, E. A., K. L. Kindle, N. Davidson, and A. Sodja. 1980. The actin genes of *Drosophila*: a dispersed multigene family. *Cell* 19:365-378.
20. Fyrberg, E. A., J. W. Mahaffey, B. J. Bond, and N. Davidson. 1983. Transcripts of the six *Drosophila* actin genes accumulate in a stage and tissue specific manner. *Cell* 33:115-123.
21. Gerlinger, P., A. Krust, M. LeMeur, F. Perrin, M. Cochet, F. Gannon, D. Dupret, and P. Chambon. 1982. Multiple initiation and polyadenylation sites for the chicken ovomucoid transcript.

- tion unit. *J. Mol. Biol.* **162**:345–364.
22. Ghosh, P., V. Reddy, J. Swinscoe, P. Choudary, P. Lebowitz, and S. Weissman. 1978. The 5' terminal leader sequences of late 16S mRNA from cells infected with simian virus 40. *J. Biol. Chem.* **253**:3643–3647.
 23. Ghosh, P., V. Reddy, J. Swinscoe, P. Lebowitz, and S. Weissman. 1978. Heterogeneity and 5'-terminal structures of the late RNAs of simian virus 40. *J. Mol. Biol.* **126**:813–846.
 24. Grez, M., H. Land, K. Glesecke, and G. Schutz. 1981. Multiple mRNAs are generated from the chicken lysozyme gene. *Cell* **25**:743–752.
 25. Hagenbuchle, O., U. Schibler, S. Petruccio, G. C. Van Tuyle, and P. K. Wellauer. 1985. Expression of mouse *Amy-2^a* alpha-amylase genes is regulated by strong pancreas-specific promoters. *J. Mol. Biol.* **185**:285–293.
 26. Hagenbuchle, O., M. Tosi, U. Schibler, R. Bovey, P. K. Wellauer, and R. A. Young. 1981. Mouse liver and salivary gland alpha-amylase mRNAs differ only in 5' nontranslated sequences. *Nature (London)* **289**:643–646.
 27. Hanauer, A., M. Levin, R. Hellig, D. Daegelen, A. Kahn, and J. L. Mandel. 1983. Isolation and characterization of cDNA clones for human skeletal muscle alpha actin. *Nucleic Acids Res.* **11**:3503–3516.
 28. Hellig, R., F. Perrin, F. Gannon, J. L. Mandel, and P. Chambon. 1980. The ovalbumin gene family: structure of the X gene and evolution of duplicated split genes. *Cell* **20**:625–637.
 29. Hunkapiller, M., S. Kent, M. Caruthers, W. Dreyer, J. Firca, C. Giffin, S. Horvath, T. Hunkapiller, P. Tempst, and L. Hood. 1984. A microchemical facility for the analysis and synthesis of genes and proteins. *Nature (London)* **310**:105–111.
 30. Lai, E. C., D. R. Roop, M. Tsai, S. L. C. Woo, and B. O'Malley. 1982. Heterogeneous initiation regions for transcription of the chicken ovomucoid gene. *Nucleic Acids Res.* **10**:5553–5567.
 31. Lee, M. G., S. A. Lewis, C. D. Wilde, and N. J. Cowan. 1983. Evolutionary history of a multigene family: an expressed human beta-tubulin gene and three processed pseudogenes. *Cell* **33**:477–487.
 32. LeMeur, M., N. Glanville, J. L. Mandel, P. Gerlinger, R. Palmiter, and P. Chambon. 1981. The ovalbumin gene family: hormonal control of X and Y gene transcription and mRNA accumulation. *Cell* **23**:561–571.
 33. LeMeur, M. A., B. Galliot, and P. Gerlinger. 1984. Termination of the ovalbumin gene transcription. *EMBO J.* **3**:2779–2786.
 34. Malek, L. T., W. H. Eschenfeldt, T. W. Munns, and R. E. Rhoads. 1981. Heterogeneity of the 5' terminus of hen ovalbumin messenger ribonucleic acid. *Nucleic Acids Res.* **9**:1657–1673.
 35. Maxam, A., and W. Gilbert. 1980. Sequencing end-labelled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499–560.
 36. McDonnell, M. W., M. N. Simon, and F. W. Studier. 1977. Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline agarose gels. *J. Mol. Biol.* **110**:119–146.
 37. Milcarek, C., and B. Hall. 1985. Cell-specific expression of secreted versus membrane forms of immunoglobulin gamma 2b mRNA involves selective use of alternate polyadenylation sites. *Mol. Cell. Biol.* **5**:2514–2520.
 38. Nabeshima, Y., Y. Fujii-Kuriyama, M. Muramatsu, and K. Ogata. 1984. Alternative transcription and two modes of splicing result in two myosin light chains from one gene. *Nature (London)* **308**:333–338.
 39. Nevins, J. R., and M. C. Wilson. 1981. Regulation of adenovirus-2 gene expression at the level of transcriptional termination and RNA processing. *Nature (London)* **290**:113–118.
 40. Parker, C. S., and J. Topol. 1984. A *Drosophila* RNA polymerase II transcription factor contains a promoter-region-specific DNA-binding activity. *Cell* **36**:357–369.
 41. Parker, V. P., S. Falkenthal, and N. Davidson. 1985. Characterization of the myosin light-chain-2 gene of *Drosophila melanogaster*. *Mol. Cell. Biol.* **5**:3058–3068.
 42. Parnes, J. R., R. R. Robinson, and J. G. Seidman. 1983. Multiple mRNA species with distinct 3' termini are transcribed from the beta₂-microglobulin gene. *Nature (London)* **302**:449–452.
 43. Perlman, D., H. O. Halvorson, and L. E. Cannon. 1982. Presecretory and cytoplasmic invertase polypeptides encoded by distinct mRNAs derived from the same structural gene differ by a signal sequence. *Proc. Natl. Acad. Sci. USA* **79**:781–785.
 44. Prochownik, E. V. 1985. Relationship between an enhancer element in the human antithrombin III gene and an immunoglobulin light-chain enhancer. *Nature (London)* **316**:845–848.
 45. Reddy, V., B. Thimmappaya, R. Dhar, K. Subramanian, S. Zain J. Pan, P. Ghosh, M. Celma, and S. Weissman. 1978. The genome of simian virus 40. *Science* **200**:494–500.
 46. Sanchez, F., S. L. Tobin, U. Rdest, E. Zulauf, and B. J. McCarthy. 1983. Two *Drosophila* actin genes in detail. Gene structure, protein structure, and transcription during development. *J. Mol. Biol.* **163**:533–551.
 47. Sasavage, N. L., M. Smith, S. Gillam, R. P. Woychik, and F. M. Rothman. 1982. Variation in the polyadenylation site of bovine prolactin mRNA. *Proc. Natl. Acad. Sci. USA* **79**:223–227.
 48. Schibler, U., O. Hagenbuchle, P. K. Wellauer, and A. C. Pittet. 1983. Two promoters of different strengths control the transcription of the mouse alpha-amylase gene *Amy-1^a* in the parotid gland and the liver. *Cell* **33**:501–508.
 49. Setzer, D. R., M. McGrogan, J. H. Nunberg, and R. T. Schimke. 1980. Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNAs in mouse cells. *Cell* **22**:361–370.
 50. Setzer, D. R., M. McGrogan, and R. T. Schimke. 1982. Nucleotide sequence surrounding multiple polyadenylation sites in the mouse dihydrofolate reductase gene. *J. Biol. Chem.* **257**:5143–5147.
 51. Snyder, M., J. Hirsh, and N. Davidson. 1981. The cuticle genes of *Drosophila*: a developmentally regulated gene cluster. *Cell* **19**:365–378.
 52. Thomas, P. S. 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci. USA* **77**:5201–5205.
 53. Tobin, S. L., E. Zulauf, F. Sanchez, E. A. Craig, and B. J. McCarthy. 1980. Multiple actin-related sequences in the *Drosophila melanogaster* genome. *Cell* **19**:121–131.
 54. Tosi, M., R. A. Young, O. Hagenbuchle, and U. Schibler. 1981. Multiple polyadenylation sites in a mouse alpha-amylase gene. *Nucleic Acids Res.* **9**:2313–2323.
 55. Unterman, R. D., K. R. Lynch, H. L. Nakhasi, K. P. Dolan, J. W. Hamilton, D. V. Cohn, and P. Feigelson. 1981. Cloning and sequence of several $\alpha_2\mu$ -globulin cDNAs. *Proc. Natl. Acad. Sci. USA* **78**:3478–3482.
 56. Wickens, M., and P. Stephenson. 1984. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent mRNA 3' end formation. *Science* **226**:1045–1051.
 57. Young, R. A., O. Hagenbuchle, and U. Schibler. 1981. A single mouse alpha-amylase gene specifies two different tissue-specific mRNAs. *Cell* **23**:451–458.
 58. Zehner, Z. E., and B. M. Peterson. 1983. Characterization of the chicken vimentin gene: single copy gene producing multiple mRNAs. *Biochemistry* **80**:911–915.